PharmaSUG 2019 - Paper AD-234

Camouflage your Clinical Trial with Machine Learning and AI

Ajith Baby Sadasivan, Genpro Life Sciences, Thiruvananthapuram, India;

Limna Salim, Genpro Life Sciences, Thiruvananthapuram, India;

Akhil Vijayan, Genpro Life Sciences, Thiruvananthapuram, India;

Bhavya K, Clap Research, Thiruvananthapuram, India

ABSTRACT

Pharmaceutical companies are currently putting a lot of focus on implementing significant changes in R&D strategies, through constant review of the scientific publications. Transparency into current research and sharing of research findings and patient level data for further research becomes crucial. In July 2018, the US Food and Drug Administration published a guidance which facilitates the use of Electronic Health Record Data in clinical investigations. Access to clinical study reports and EHR (through FHIR) has been granted to the research community in order to accelerate collaborative research.

Anonymization and Pseudonymization have been the topic of discussion since the introduction of the General Data Protection Regulation. GDPR acknowledges the privacy augmentation provided by these techniques. The document provides exceptions to many of the tough provisions of the regulation, when personal data is anonymised. If the trail that connects personal data to an identifiable person is lost, then, data managers are permitted to use, process and publish personal information in all possible methods that enable collaborative research

So far the industry has taken an approach of engaging external agencies to process CSR documents. Electronic redaction (which is the equivalent of drawing a thick black line through patient information) is an option, according to the new guidance. Some of the companies have even started investing heavily on the source data anonymisation and generating the documents again for submission. As part of this process, we need to protect a patient's identity - which could be open to discovery based on the type of study they took part in, their age, race and demographic, and when they attended hospital or clinic, among other bits of information in their clinical data record. These manual processes usually involve challenges in terms of cost of processing, quality and turn around times.

This paper explores the possibilities of developing a dynamic machine learning framework using spaCy in compliance with the EMA Policy 0070 for easy and effective anonymization of clinical reports by generating a named entity recognition (NER) model which will automatically identify the variables that needs to be anonymised from a report. For eg, the system will be able to identify all the patient IDs, Site IDs from a clinical study report and then run further anonymisation techniques on top of their regions. The objective of this framework is to largely automate the anonymisation process and later play a major role in quality controlling the documents for compliance by reconciling it with source data.

The system will ingest PDF reports or source trial data, use the NER model to identify the entities that needs to be anonymised, perform anonymization algorithms and then regenerate the PDF's with anonymised data which will be ready for compliant submissions.

The biggest challenge in this process is to identify the personal or quasi identifiers to be anonymized from reports and to mask it such that the original demeanor is not altered. The paper describes methods to overcome this with the help of AI and ML methodologies and gives the user the authority to approve the masking of required ID's proposed by the tool itself after proper training.

INTRODUCTION

Life sciences industry is now faced with a market and regulatory expectation to be more transparent with the clinical trials. Patients who participated in the study feels that they are entitled to know the findings along with the general public who would like to monitor the long-term outcomes of the research. Most of the large pharmaceutical companies have started investing in processes, technologies and platforms for making patient level data available for public scrutiny. This will enable collaborative research across life science community and better preparedness to face regulatory audits and interventions

European Medicines Agency (EMA) had proactively initiated steps in this direction. They had formulated a new policy for clinical trial reporting (EMA Policy 0070) in 2014. The agency has also issued a 91 page implementation guidance document for life sciences organisations to comply with the new requirements. Marketing authorisation process already requires the submission of clinical study reports (CSRs). The new regulation dictates that, within 60 days of an authorisation decision (positive or otherwise), the CSRs must be made available in a format that removes any risk of a subject's identity being compromised.

Majority of life science companies has been investing into tools that will help them anonymise the data that is part of the report. But the pace at which technology is being adopted to do that has not been a swift process. Even though the guidance issued initially prioritises clinical study reports as the primary target for patient anonymisation, the regulation is going to now cover all reports related to clinical studies. Hence it becomes extremely important for organisations to adopt a fast, cost effective, scalable and sustainable solution which can handle large volumes of data with accuracy over a short span of time.

One of the most comprehensive methods to approach patient anonymisation in clinical trial reports is to start with the patient-level data. If you can get your patient data anonymised then the pipeline that follows will be fool proof. This will save a lot of time and mitigate risk in the longer term and probably is the only way in which patient data can be processed consistently. This consistency is critical in ensuring that study findings retain their scientific meaning and value.

It is also important to use algorithms consistently between documents and data for patient anonymisation. Else, it becomes increasingly difficult to reconcile the anomalies as researchers later need to perform further cross referencing and analysis. This will be unnecessary complexity and more work for companies down the line as they try and address numerous follow-on questions once clinical trial findings are in the public domain.

Hence it becomes increasingly important to create a framework which will help companies to ingest document, data along with the variables to be anonymised and then automate the process of identifying the variables from data and documents, apply anonymization algorithms and generate an anonymisation report for further audits. This paper puts together a technology pipeline to automate the entire process by using AI and NLP algorithms along with a PDF parser and basic data management interface.

DATA ANONYMIZATION

Data anonymization provides a mechanism to manage the tension between safeguarding personal privacy and maximising the utility of data. If the data is anonymized, then it is no longer considered as personal data and can be shared and reused. This helps to maximise the benefits of sharing of healthcare data, pooling and integration with other datasets.

Anonymisation Methods

Some of the methods that are used for data anonymisation is listed below:

1. Record Suppression
2. Cell Suppression
3. Randomization
4. Shuffling
5. Creating Pseudonyms or Surrogate

6. Sub Sampling
7. Aggregation/ Generalization
8. Adding Noise
9. Character Scrambling
10. Character Masking
11. Truncation
12. Encoding
13. Blurring
14. Masking
15. Perturbation
16. Redaction

Documents

Although EMEA has started with clinical study reports as its target, the published policy document indicates it is only a matter of time before all of the patient-level data behind those reports will need to be given the same treatment. This means, all reports, publications, narratives posters and other documents which might contain any of the 'variables of interest' will have to go through the process of data anonymisation to meet the compliance requirements.

Variables of Interest

Some of the variables that need to be anonymised as part of the process could include the following. Please note that this is not a comprehensive list but an indicative list.

Patient Level data

- Patient identification number (Screening number and/or randomization number)
- Site identification number (in case of multiple sites)
- External ID's, CRO/TPO IDs,
- Dates: Visit dates, Randomization date, Birth date, Adverse event start and stop date, serious adverse event start and stop date, Death date
- Age information (in case of elderly subjects >89 years of age or any other study specific age criteria)

Data to be Redacted

- Individual Level Information (Subjects in the Trial)
- Personal information: Name, initials, email, phone number, signature, full address including country (in case of less number of patients)
- Geographic information such as place of work, trial site location, addresses, zip codes, etc.
- Biometric identifiers including such as magnetic resonance imaging outputs with any patient identifier, hand/voice prints, facial images that may lead to patient identification etc.

Individual Level Information (Site Staff)

- Personal information: Name, initials, email, phone number, signature, full address
- Investigator's and Trial Site Personnel's' Curriculum Vitae

Data to be Retained

- Academic qualifications (e.g. MD, PhD)
- Committee names and address including countries
- Name of the Principal investigator and Sponsor Company's address, own research lab name, addresses, country
- Study roles (e.g. Principal Investigator, Statistician)
- CRO's/Third party research lab name, address, country
- Dates not related to study participants (e.g.patient visit dates, sample collection dates)

CURRENT PROCESS

As the EMA's guidance are so new, the industry hasn't had much time to follow it. So, what the companies are doing currently? Some firms have taken the easy way out- Engaging external agencies to process CSR documents. Based on the new guidance, Electronic redaction- drawing thick black lines through patient information is not an advisable solution.

Considering the time limits, some of the Life science companies have also developed tools to anonymize the clinical trial data. Most of these tools are working only on a pre-identified set of variables or labels with fixed algorithms to anonymize the data. So, anonymization of documents such as Patient Narratives or Clinical Study Reports becomes challenging. With the help of SAS or some other programming languages, the pre-identified variables like Patient Id, Age, Race and other demographic information are Masked/Grouped/Removed from the data- which is not an ideal solution to the problem.

PROPOSED SYSTEM

The paper proposes a data anonymization framework which can input, PDF documents, structured data and any associated documents like a clinical protocol and SAP if required. User will also be able to specify the variables that they want to get anonymised. The system will parse the PDF documents, use the pretrained NER model to extract the variables that need to be anonymised, use an anonymisation engine to perform the anonymisation and use the PDF generator to regenerate the documents. The system will also give an anonymisation analytics report which will give an overview of the list of parameters identified and anonymised with in each document.

This system will help pharmaceutical companies to semi automate the process of anonymising their clinical reports. When a more robust source data anonymisation system is put into place, this framework will help companies to perform QC on their anonymisation by extracting data out of the documents and reconciling with the source data.

Figure 1 is the design of the system.

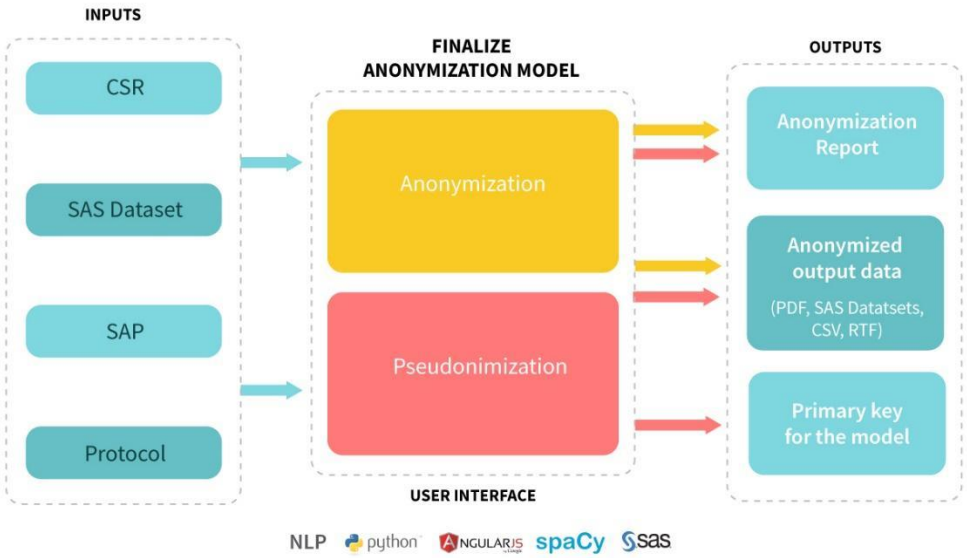

Figure 1. Design of the system

FEATURES

PDF PARSER

Most of the study reports and narratives comes in the form of PDF documents. This framework will incorporate a PDF parser which will take a PDF document as input and provide the contents of the document in a neat structure. The structure will be similar to HTML which will help you identify the text, tables and images in the PDF and then reconstruct the document in a structured fashion.

Named Entity Recognition

Named Entity Recognition also known as NER is a main task in information extraction. Entity extraction classifies named entities that are present in a text into pre-defined categories like “individuals”, “companies”, “places”, “organization”, “cities”, “dates”, “product terminologies” etc.

Example : ”Leland Stanford Junior University is a private research university in California.“

Named Entities : Leland Stanford Junior University -> ORGANIZATION

California -> LOCATION

There exist different ways to identify and extract named entities from text.

Dictionary based

The simplest method is to keep a dictionary for all possible named entities and doing a lookup for every input sentence in the dataset for named entities. The method benefits from computational cost, time, etc. But there is no understanding about the context where the entities are identified.

Grammar based

In grammar-based method expertly curated grammars are generated to identify and extract entities from unstructured text. Grammar based techniques require a lot of man-hours from experienced computational linguistics to create hand-crafted grammar-based solutions which will have good precision, but a lower recall. These techniques have become less popular since the arrival of better machine learning based techniques.

Machine Learning based

Supervised machine learning models learn to make predictions by training on example inputs and their expected outputs, and can be used to replace human curated rules. Hidden Markov Models (HMM), Support Vector Machines (SVM), Conditional Random Fields (CRF), and decision trees were common machine learning systems for NER.

Datasets available for NER training

Annotated Corpus for Named Entity Recognition: Corpus (CoNLL 2002)

N3 - A collection of datasets for NE

Annotated Corpus for Named Entity Recognition by Anton Dmitriev

The Groningen Meaning Bank (GMB Dataset from the OKE Challenge 2016

spaCy NER Model

Being a free and an open-source library, spaCy has made advanced Natural Language Processing (NLP) much simpler in Python.spaCy provides an exceptionally efficient statistical system for named entity recognition in python, which can assign labels to groups of tokens which are contiguous. It provides a default model which can recognize a wide range of named or numerical entities, which include company-name, location, organization, product-name, etc to name a few. Apart from these default entities, spaCy enables the addition of arbitrary classes to the entity-recognition model, by training the model to update it with newer trained examples.

spaCy's models are statistical and every "decision" they make – for example, which part-of-speech tag to assign, or whether a word is a named entity – is a prediction. This prediction is based on the examples the model has seen during training.

The model will be shown to unseen examples and it will make some predictions. Since we know the actual label, we compare the output model predicted and the actual label in the dataset and this is known as the error in training. This error will be back propagated to the network and the model eventually learn to minimize the error. For the model to be generic to new examples, the training data should be always a representative of the data we are going to test with. We used the existing Spacy model to make it learn new entities. Annotated the text extracted from the pdf documents and train the model for custom entities.

Training Steps

The system uses Named Entity Recognition (NER) model to identify the entities within the unstructured text. Named Entity Recognition is an information extraction sub task to find and classify the named entity mentioned in structured text. Some of the entities that we need to extract from the CSR will include, Patient ID, Site ID etc.

As part of the training, we will use sentences from different types of CSR which mentions patient ID and use them to train a NER model. A total of 3 CSRs were processed out of which 100 sentences with subject id mentions were extracted as a training dataset. Using this training dataset, the NER model was trained to predict patient ID in the test documents (1 in this case).

Figure 2 explains the process of training.

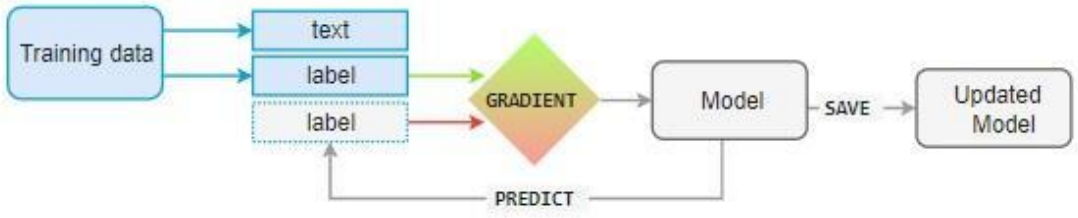

Figure 2. System Building

Suppose we have two new patterns of data:

1. SAE occurred for subject 101-007 on 2018-02-14
2. Patient 040-011 is a 69-year-old white female who was first diagnosed with multiple myeloma (MM; ISS Stage 3) in November 1999.

The below snippet shows the format in which we feed input for training the NER model. The first sample explains that the string starting from 24 to 30 should be considered as – “PATIENT_ID”. Similarly, the second sample shows how the entities like AGE, SEX and DATE are also added to the NER Model. Once the model is updated with these definitions, the system will be capable of understanding any text of this format.

Figure 3 shows the format in which we feed input for training the NER model.

```
(
    "SAE occured for subject 101-007 on 2018-02-14",
        [
            (24, 30, 'PATIENT_ID')
        ]
)

(
"Patient 040-011 is a 69-year-old white female who was first diagnosed with multiple myeloma (MM; ISS
Stage 3) in November 1999",
    [
        (8, 14, 'PATIENT_ID'),
        (21, 22, 'AGE'),
        (39, 44, 'SEX'),
        (113, 125, 'DATE')
    ]
)
```

Figure 3. Training Data Examples

The system will use the training data that has been collected across multiple documents and then use that to train the NER model. As can be seen, we will train the model until the loss (error) will fall within reasonable limits.

Figure 4 shows the SpaCy NER training loss.

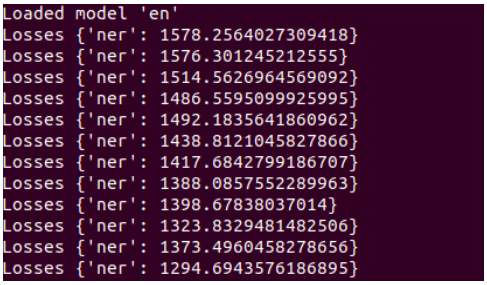

Figure 4. SpaCy NER Training loss

Once the model is trained, we will be able to use it to automatically identify the variables within the document. As a test, we have trained a NER model which will identify and extract subject IDs from the clinical reports.

Figure 5 gives the screenshots that will depict the model in production.

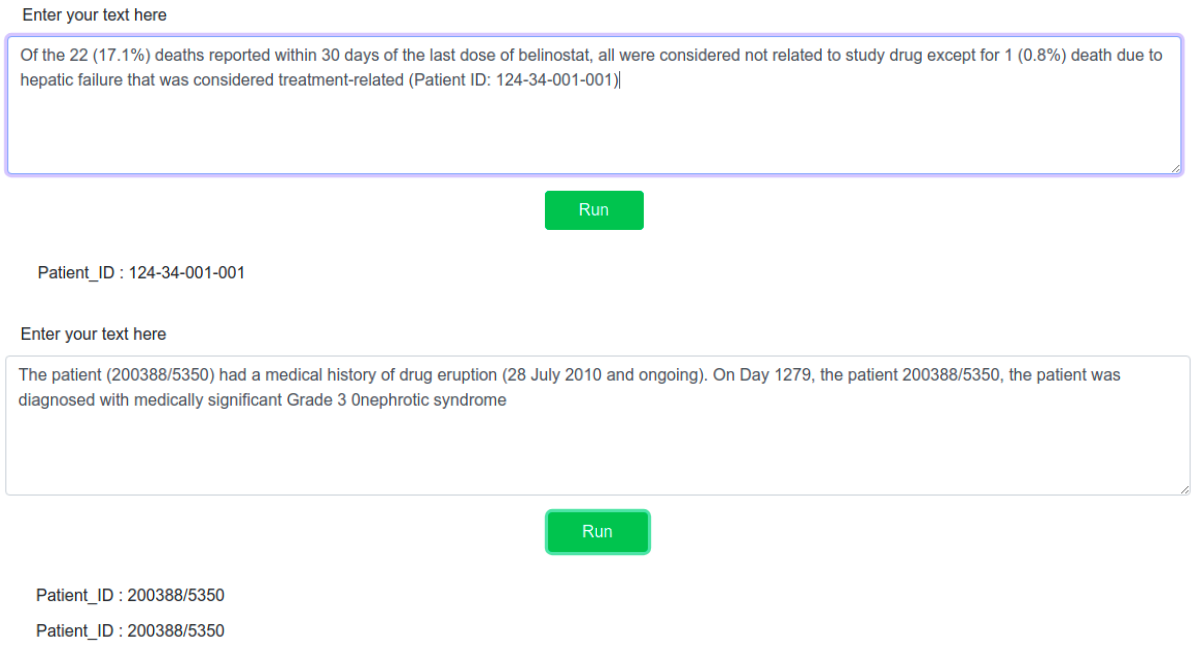

Figure 5. NER Extraction Results

Anonymisation Engine

Once the specified variables are extracted, the system will provide an anonymisation engine which will perform the algorithms in order to anonymise/psuedonymise the data. Such transformed data will be stored in a database

PDF Generator

The document generator will be the module which will use the anonymised data and the tody structure to replace the variables with decrypted values. Once the replacement is done, this module will further generate the PDF again.

Anonymisation Report

This module will serve as a dashboard where analytics of all anonymisation performed on a document set will be displayed to the user. The user will be able to drill down further and cross check whether the system has identified the right variables and the anonymisation performed properly. Some of the data points that this module presents will include the following:

- Total number of documents processed
- Total variables identified
- Old Value/ Decrypted Value with source document
- Time taken per document for processing

Potential Users

This framework could be used by the following teams from a clinical research team

- Regulatory and Submissions
- Data Management
- Clinical Data Management
- Medical Writing
- Quality Control

Way Forward

Currently Genpro is looking for early adopters who will be able to work alongside our machine learning team and supply enough clinical reports for us to train a NER model which is ready for production. Our team would like to pursue this as a commercial engagement, once we have established success in automatically anonymising the clinical reports.

CONCLUSION

Artificial Intelligence and Natural Language Processing is going to revolutionize the way in which document management and processing is done in every vertical. In Clinical trials, EMA gudiances on data protection has provided an excellent opportunity for us to put these powerful technologies to use. We hope this document will serve as a guiding document for companies and teams who would like to automate their clinical report anonymisation process.

REFERENCES

Data anonymisation - a key enabler for clinical data sharing - https://www.ema.europa.eu/en/documents/report/report-data-anonymisation-key-enabler-clinical-data-sharing_en.pdf

Nested Named Entity Recognition - https://nlp.stanford.edu/pubs/nested-ner.pdf

Natural Language Toolkit - https://www.nltk.org/

Named Entity Recognition with NLTK and SpaCy - https://towardsdatascience.com/named-entity-recognition-with-nltk-and-spacy-8c4a7d88e7da

Data Anonymization Algorithms Based on K-Anonymity - https://iopscience.iop.org/article/10.1088/1757-899X/225/1/012279/pdf

https://spacy.io/

ACKNOWLEDGMENTS

The content, ideas and recommendations presented in this paper are all developed from experiences in our career. These experiences come through previous companies, various industry leaders, colleagues, mentors, conferences and direct experience.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the primary author at:

Author Name: Ajith Baby Sadasivan
Company : Genpro Life Sciences India Private Ltd
Work Phone : 0471-2700151/2700152
Email : ajith.nair@genproindia.com
Web : www.genproindia.com

Author Name: Bhavya K
Company : Clap Research
Work Phone : 0471-2700151/2700152
Email : bhavya.k@clapresearch.com
Web : www.clapresearch.com

Author Name: Limna Salim
Company : Genpro Life Sciences India Private Ltd
Work Phone : 0471-2700151/2700152
Email : limna.salim@genproindia.com
Web : www.genproindia.com

Author Name: Akhil Vijayan
Company : Genpro Life Sciences India Private Ltd
Work Phone : 0471-2700151/2700152
Email : akhil.vijayan@genproindia.com
Web : www.genproindia.com